ELISA, CLIPS and LIA NIST 2003 segmentation

ELISA consortium = CLIPS and LIA for RT

CLIPS Grenoble (Fr)
Daniel Moraru
Laurent Besacier

LIA Avignon (Fr)
Sylvain Meignier
Corinne Fredouille
Jean-François Bonastre

Outline

- Segmentation component overview
- Acoustic segmentation
- Speaker segmentation
  - CLIPS approach
  - LIA approach
- Speaker re-segmentation
- ELISA collaboration
  - Merged system
  - Piped system
- Results

Segmentation component overview

- Based on acoustic pre-segmentation
- Speaker segmentation
  - LIA
    - based on HMM
    - Re-segmentation at the end
  - CLIPS
    - BIC detector
    - based on hierarchical clustering
  - ELISA
    - Piped
      - Acoustic pre-seg. → CLIPS speaker Seg → LIA re-segmentation
    - Merged
      - Merging of 4 speaker segmentation systems → LIA re-segmentation

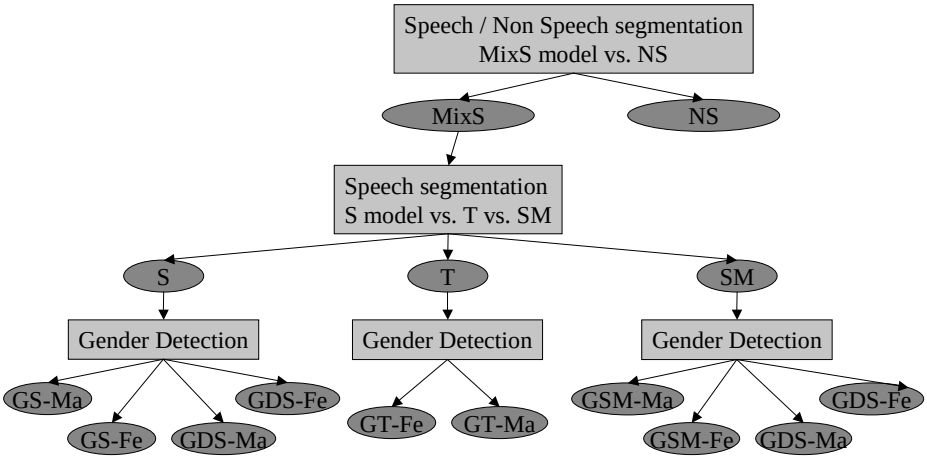
Acoustic pre-segmentation

- Objectives
  - To provide an acoustic pre-segmentation to speaker segmentation phase based on :
    - Speech / Non speech detection
    - Wide band / Narrow band detection (~ studio/telephone)
    - Gender detection
- Approach
  - GMM model based
  - Viterbi decoding
  - Hierarchical segmentation

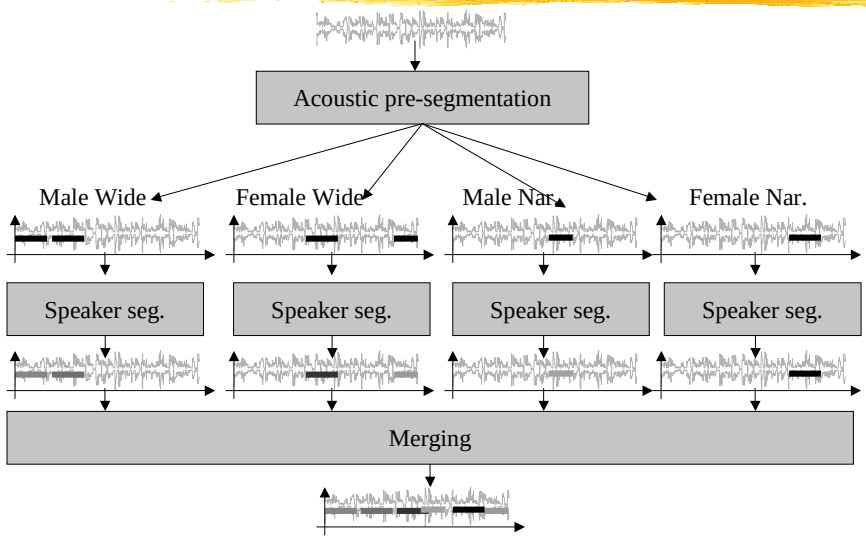
Acoustic pre-segmentation

- Feature vector
  - 12 MFCC + E + Δ + ΔE + ΔΔ + ΔΔE = 39 coef.
  - No CMS, Windows = 25ms, delay=10ms
- Acoustic modeling = GMM diagonal
  - Non speech : NS = 1 component
  - Wide speech
    - S : Gender indep. speech model = 1024 comp. *(BN condition : F0 & F1)*
    - SM : Gender indep. speech over music model = 1024 comp. *(F3)*
    - GS-Ma & GS-Fe : Gender dep. speech models = 2x128 comp. *(F0 & F1)*
    - GSM-Ma & GSM-Fe : Gender dep. speech over music models = 2x128 comp. *(F3)*
    - GDS-Ma & GDS-Fe : Gender dep. degraded speech models = 2x128 comp. *(F4)*
    - MixS : merging of GS-Ma + GS-Fe + GDS-Ma + GDS-Fa = 512 comp.
  - Narrow speech
    - T : Gender indep. speech model = 1024 comp. *(Telephone from F2)*
    - GT-Ma & GT-Fe : Gender dep. speech models = 2x128 comp. *(Telephone from F2)*
  - Learned on a subset of BN 96 corpus
- Acoustic condition bi-gram probability learned on BN 96 corpus

Acoustic pre-segmentation : hierarchical approach



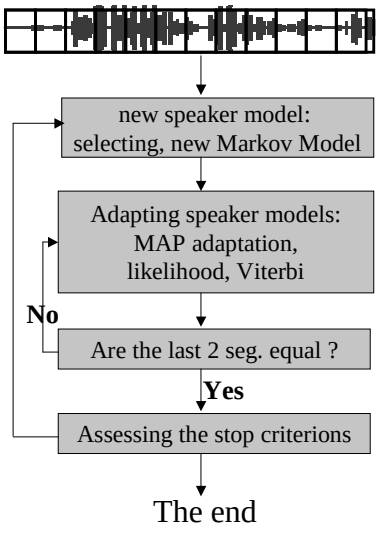
Speaker segmentation



Speaker segmentation LIA vs CLIPS

LIA system	CLIPS system
Parameterization 20LFCC+E, no CMS, no bandlimiting	Parameterization 16MFCC +E, no CMS, no bandlimiting
"Segmentation *a priori*" Every 0.3 s	Segmentation BIC + Acoustic Segmentation Using 1.75 sec windows
Clustering **descendant** based on HMM one state = one speaker uses background model LIA-MAP speaker adaptation	Hierarchical Clustering **ascendant** uses GLR as distance uses a background model MAP speaker adaptation

Speaker segmentation: LIA

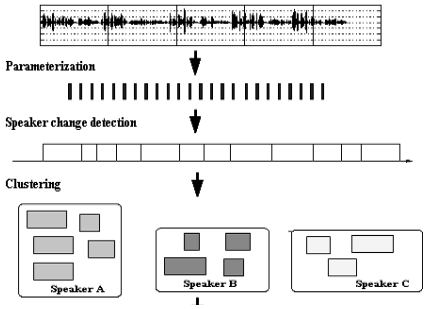


- Initialization:
  - Markov Model with one state (L1)
  - Segmentation is composed of one segment labeled L1.
- New Speaker L2
  - built from the 3s block with maximum likelihood computed on L1
- Iterative adaptation of the MM
  - Adaptation and Viterbi decoding
  - If two consecutive segmentations are equal, we stop
- Speaker model validation
  - The last speaker segment duration is less than 4s
  - The previous speaker is removed if the new one is longer
- We stop the segmentation if:
  - no more 3s blocks labeled L1

Speaker segmentation: LIA

- Background model (UBM)
  - Subset BN 96 : F0 F1 F2 condition (wide and narrow)
    - No gender dep.
    - No band dep.
- LIA baseline (primary)
  - Speaker adaptation : LIA MAP (same as 1-speaker task)
  - Mean only
  - Dependent of model weights
- LIA MAP2
  - Speaker adaptation : linear MAP
    - Mean only

Speaker Segmentation: CLIPS



- One unique system
- Speaker change detection
  - BIC distance
  - 1.75 sec adjacent windows
  - mono-gaussian models with diagonal covariance matrix
  - Acoustic Segmentation
- Clustering
  - diagonal 32 GMM background model learned on the entire file
  - MAP adaptation (means)
- Estimate the number of speaker
  - uses BIC maximization

Estimate N : CLIPS

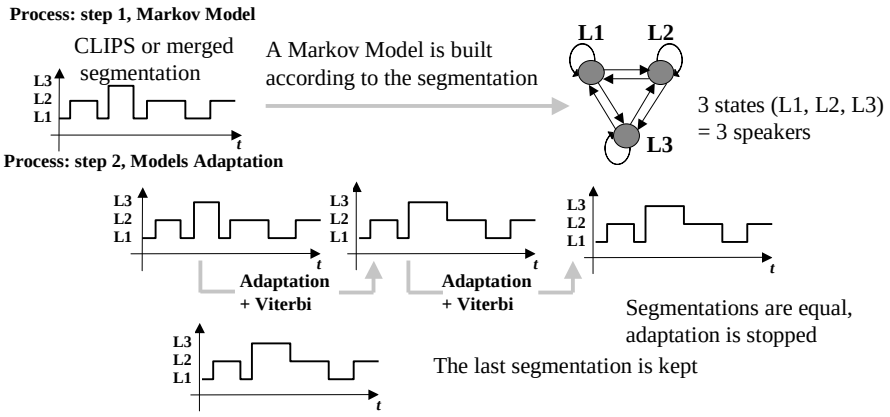
- The segmentation is done independently for each class given by the acoustic pre-segmentation
- Estimate the number of speakers using the BIC criterion
- Limit the number of speakers (N) between 1 and 25
- Select N that maximizes the BIC criterion

$$BIC\,(M)=\log L(X;M)-\lambda\frac{m}{2}N_{Sp}\log N_X$$

- Algorithm developed with the help of I. Magrin-Chagnolleau at the DDL Laboratory in Lyon

Speaker re-segmentation

- LIA re-segmentation
  - Based on the HMM speaker detection phase
  - Speaker adaptation = MAP-MIT, r=16

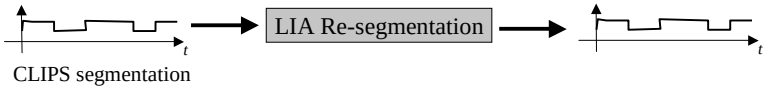


ELISA collaboration

- Hybrid systems
  - piped system
    - CLIPS segmentation re-segmented by the LIA system
  - Merged system
    - Merging of 4 segmentations before LIA re-segmentation

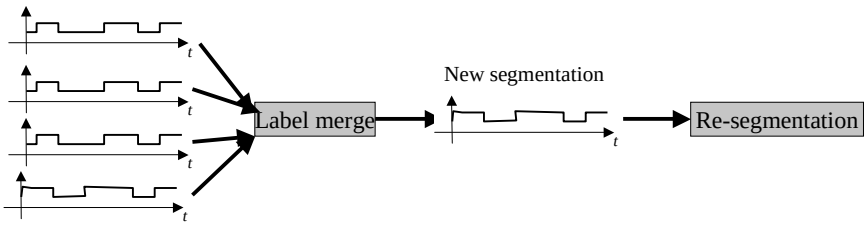
ELISA collaboration: pipe

- Uses the results of one system to initialize the other
  - CLIPS segmentation piped in LIA re-segmentation system



ELISA collaboration: merge

- Merging of 4 results resulting from LIA and CLIPS segmentations
  - Use both segmentations
    - LIA baseline (primary)
    - LIA MAP2
    - CLIPS (primary)
    - CLIPS piped LIA
- Merging: labels by frame are merged
  - Ex:
    - T_0: Sys1="S1", Sys2="T4", Sys3="S1", Sys4="F1" → "S1T4S1F1"
    - T_1: Sys1="**S2**", Sys2="T4", Sys3="S1", Sys4="F1" → "**S2**T4S1F1"



Results

CLIPS (Primary)	19.25%
LIA MAP-LIA (Primary) LIA MAP-linear	16.90% 24.71%
ELISA Merge (Primary) ELISA Pipe	14.24% 12.88%

- The collaboration systems (ELISA) improved performance of starting systems
- The merged system gives the possibility to use multiple segmentation systems

- There is still a lot to gain of the acoustic pre-segmentation and the estimation of the number of speakers

References and contacts

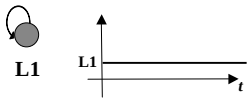
- ICASSP 2003, "The ELISA Consortium Approaches in Speaker Segmentation during The NIST 2002 Speaker Recognition Evaluation", D. Moraru, S. Meignier, L. Besacier, J.-F. Bonastre, I. Magrin-Chagnolleau

- Contacts:
  - daniel.moraru@imag.fr
  - sylvain.meignier@lia.univ-avignon.fr

Segmentation system: indexing process (stage 1)

Stage 1: adding speaker L1

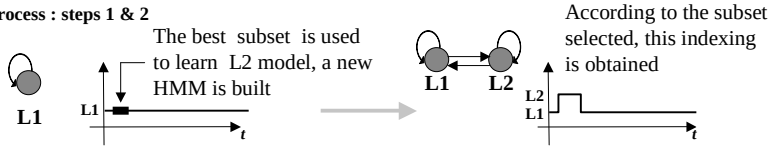
Process initialization



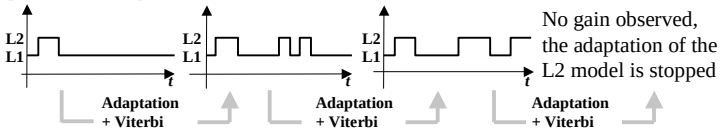
Segmentation system: indexing process (stage 2)

Stage 2: adding speaker L2

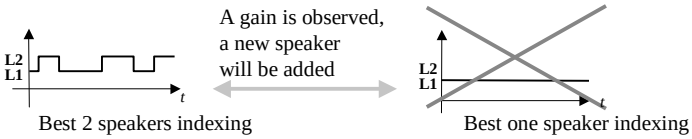
Process : steps 1 & 2



Process : step 3, Models Adaptation

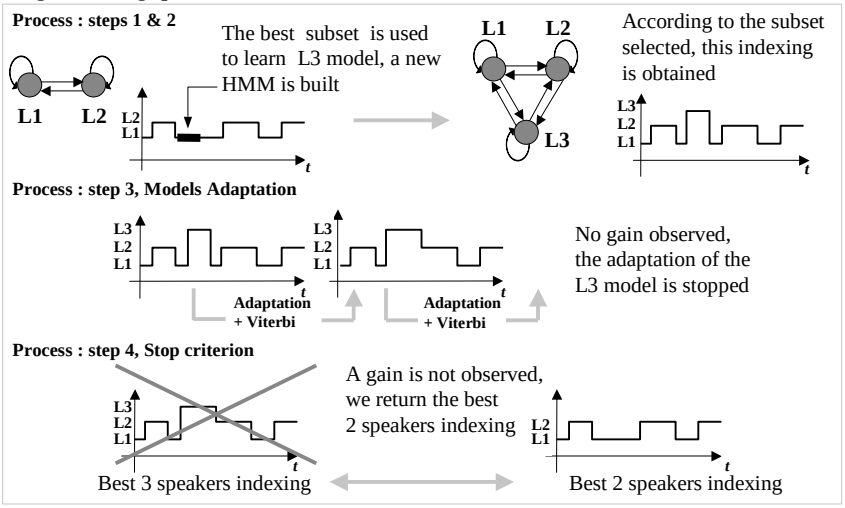


Process : step 4, Stop criterion

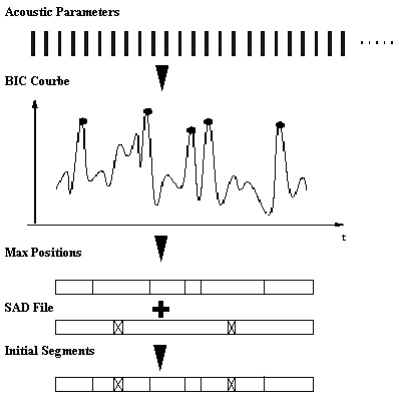


Segmentation system: indexing process (stage 3)

Stage 3: adding speaker L3



CLIPS : Speaker change detection

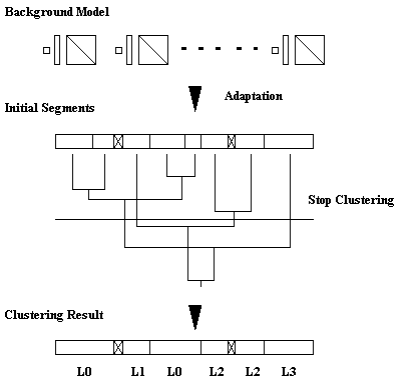


- BIC distance computed using two adjacent sliding windows

$$d_{BIC}(w_1, w_2) = \frac{p(w1/M_1)\cdot p(w2/M_2)}{p(w1, w2/M_{12})}$$

- Windows are modeled using mono-gaussian models with diagonal covariance matrix
- Look for maximums of BIC distance
- Use pre-segmentation

CLIPS : Clustering



- Train a diagonal GMM 32 background model on the entire file
- Three pass EM
- Adapt the background model on every segment
- BIC distance is computed and the two closest segments are merged
- Estimate N